



Multi-Lingual Emotion Classification using Convolutional Neural Networks

Alexander Iliev^{1,2}(✉), Ameya Mote², and Arjun Manoharan²

¹ Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
ailiev@berkeley.edu

² SRH University Berlin, Charlottenburg, Germany
3104966@stud.srh-campus-berlin.de;
3105614@stud.srh-campus-berlin.de

Abstract. Emotions play a central role in human interaction. Interestingly, different cultures have different ways to express the same emotion, and this motivates the need to study the way emotions are expressed in different parts of the world to understand these differences better. This paper aims to compare 4 emotions namely, anger, happiness, sadness, and neutral as expressed by speakers from 4 different languages (Canadian French, Italian, North American English and German - Berlin Deutsche) using modern digital signal processing methods and convolutional neural networks.

Keywords: digital culture • multimedia systems • information retrieval • audio analysis • audio analysis • speech processing • multi-culture comparison • emotion recognition • convolutional neural networks.

1 INTRODUCTION

In a new social environment, emotions are one-way people use as hint to correctly behave in front of other observers. Negative emotions like anger or sadness by the viewers tells whether the person performing the task in doing it correctly or not. Positive emotions signal the opposite of the former [1]. William James, in 1890, classified emotions into 4 basic types. They are namely: Fear, Grief, Love, and Rage. Paul Ekman, later classified emotions into 6 basic types and this classification is recognized worldwide in the following set of emotions: happiness, anger, sadness, disgust, fear and surprise.

Although there is wide consensus on the identity of these basic emotions, the way they are expressed in different parts of the world is unique to the people inhabiting those regions. In a globally connected world with constant interaction between individuals of vastly different cultural backgrounds, it becomes important to study the differences in the way emotions are expressed and document them such that, we as a collective human species can understand and work through these variabilities. Proper understanding of emotional cues forms the bedrock of a healthy and successful human interaction and thus warrants the

Bulgarian Ministry of Education and Science

need to better understand subtle differences in the variabilities of expression of these emotions. Through this paper, we seek to understand the various nuances of emotional expression using speech data as expressed by native speakers of different languages. This paper is a follow up paper to the one published by the authors at DIPP: 2020 in the relevant area of research. In this work, we have added more datasets and hence hope to make this work more expansive, while improving on the techniques used in our original paper.

2 DATASET

Our dataset consists of speech data of various actors enacting different emotions in Canadian French[2], Italian [3], North American English [4], Berlin Deutsche[?] respectively. The datasets contain audio samples of both male and female speakers. The ages of the people were random, and no particular emphasis was given to this metric. Due to unavailability of a single source, the datasets were procured from different sources, but effort was taken to standardize it in terms of recording quality and audio channel.

Table 1. Database Description.

Dataset	Male	Female	Emotions	Audio Quality	Total Sentences
Canadian French	6	6	Anger, Disgust, Fear, Happiness, Sad, Neutral	48KHz,Mono	443
EMOVO Italian	3	3	Anger, Disgust, Fear, Happiness, Sad, Neutral	48KHz,Mono	510
North American	12	12	Anger, Disgust, Fear, Happiness, Sad, Neutral,Calm	48KHz,Mono	1440
Berlin Deutsche	5	5	Anger, Disgust, Fear, Happiness, Sad, Neutral	16KHz,Mono	535

A large dataset is required to achieve good results with convolutional neural networks. As our datasets were not large enough, we used data augmentation techniques to enlarge it. Data augmentation refers to the technique of enlarging datasets by adding slight modifications to the original data and is a widely practised trick in the data science community. Since our data consists of audio files, we used the following techniques to enlarge our data [6]:

Pitch tuning: The pitch of the audio was shifted to create a different sample but with the same audio. This augmented dataset had different pitch characteristics but expressed the same emotions as of the original audio samples.

Addition of white noise: A random white noise was introduced into the audio sample. On a superficial level, humans ears distinguish these sounds as a background buzz.

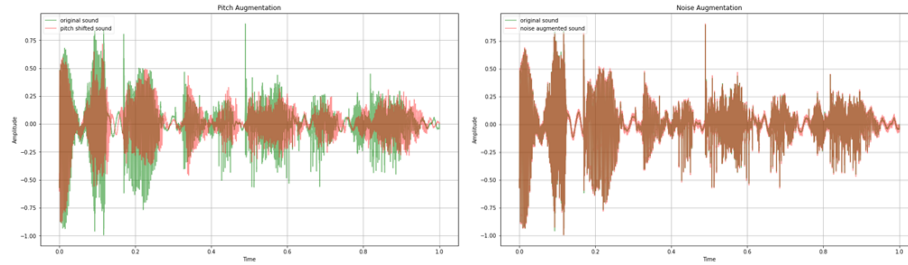


Fig.1. Original vs Pitch and Noise Shifted Sample

From Fig.1, we see that although the two signals are visibly different, but they still have the same overall characteristics. The difference introduced by the data augmentation techniques is enough to make the augmented data have different features in comparison to the their originals different without altering the characteristics we are interested it and thus we effectively enlarge our dataset.

Table 2. Data Augmentation.

Dataset	Number of samples before augmentation	Number of samples after augmentation
Canadian French	299	897
EMOVO Italian	286	858
North American	299	897
Berlin Deutsche	301	603

After enlarging the dataset, the samples needed to be transformed in such relevant information could be captured. For this purpose, we decided to use the Mel Frequency Cepstrum associated with each audio file as the feature vector representing that audio clip.

Detailing it further, the raw file in 3a, German male happy is converted into the frequency scale using the Fast Fourier Transform. The image 3b shows the audio sample in the frequency spectrum. The audio is then again converted into the Mel frequency scale as shown by the Spectrogram in image 4a. The conversion gives us the Mel Frequency Cepstral Coefficients (MFCC). These features are utilized for training the model. 4b gives us the variation of the audio utterances at different points of time, i.e., 2.5 seconds of 65 blocks.

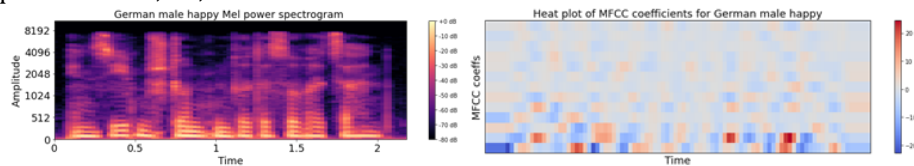


Fig.2. Figure (2a-Berlin Deutsche male happy Mel power spectrogram), Figure (2bHeat plot for MFCC coefficients for Berlin Deutsche male happy)

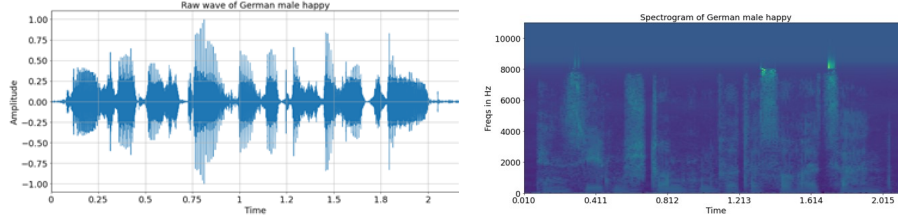


Fig.3. Figure (3a-Raw wave of Berlin Deutsche male happy), Figure (3b-Spectrogram of Berlin Deutsche male happy)

3 Model

3.1 Mel Frequency Spectrum

For achieving a good classification technique, it always becomes utmost important to have similar data available. The datasets used herein have been gathered from various sources and after applying a data augmentation are now somewhat like each other. To achieve more similarity and to gather necessary features we use MFCC [7]. It is well documented how Human ears cannot perceive about 1000 Hz. For this purpose, Mel frequency can be used via 2 filters: below 1kHz, they are spaced linearly, whereas above its logarithmic equations are used[8]. The equation for the frequency is given by,
Displayed equations are centered and set on a separate line.

$$F(m) = [1000/\log_{10}(2)] * [\log_{10}(1 + f/1000)], \quad (1)$$

F(m) stands for Mel-Frequency and f stands for Frequency in Hz.

$$MFCC_m = \sum_{k=1}^1 \log_{10}(E_k) [m(k - 1/2)/20], \quad (2)$$

Here m stands for the total number of Mel-Frequency cepstral coefficients and Ek stands the total number for the energy output of the Kth filter.

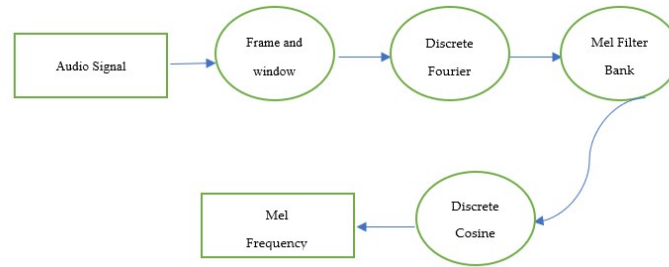


Fig.4. MFCC

3.2 Neural Network Description

To ensure that MFCC's capture the features associated with emotions, we trained a convolutional neural network and tested if our hypothesis, we needed a technique which could give us good results. For this, Convolutional Neural Network was used[9]. The 13 MFCC features chosen were deemed capable to compare the emotions across different cultures to each other. To achieve high accuracy, it is suggested to use many layers and that was incorporated.

1. Conv1d: There are in total 8 convolutional layers in our network.
2. Activation: One activation layer was associated with each of the conv1d layers. Of the, thus 8 layers, 7 of them are ReLU activation layers and the activation layer associated with the final layer is 'Softmax' layer.
3. Batch Normalisation: As the training set is fed into the network as minibatches, a batch normalization layer was implemented to normalize the inputs of each mini batch.
4. Dropout: This technique is used to remove the problem of overfitting. A few features are randomly dropped so that the model does not completely learn everything. This helps the model in learning more important features and ignore other random features. Dropout layer was added after the 2nd and the 6th layer[10]
5. MaxPooling: Maxpooling is a regularization technique and also serves as a dimension reduction technique. It combines all the important features together and passes on the most relevant patterns and helps the model can learn faster.
- 6, Flatten: Since the model requires a single vector, flatten unravels the input matrix into one long feature vector.
7. Dense: The final layer of the network was a fully connected layer with a sigmoid activation layer [11].

3.3 Train and Test Samples

The datasets for the three languages French, Italian and North American English were split into training and test sets as discussed in [12] for the four emotions:

happy, angry, sad, neutral. An extended including the additional Berlin Deutsche dataset is shown above.

Table 3. Berlin Deutsche Train and Test Samples.

Emotion	Male	Female	Train	Test
Happy	57	75	46(Male), 60(Female)	11(Male), 15(Female)
Anger	102	111	81(Male), 89(Female)	21(Male), 22(Female)
Sad	51	51	41(Male), 41(Female)	10(Male), 10(Female)
Neutral	72	84	57(Male), 67(Female)	15(Male), 17(Female)

As compared to the previous paper[12], adding another language helps in understanding how robust the CNN model is. Earlier, after comparing all 3 languages, on an average, it was found out that 71 percent of all emotions are similar to each other. While the results being good, It really did not offer much in terms of complexity. As a result, in the present work, it was decided to use another language set and compare it with other languages to understand if a higher order complexity could be achieved with the same CNN model. Thus, The Berlin Deutsche set was employed.

4 Results

Prior to using data augmentation, we had achieved an average accuracy of 53 percent which was deemed unacceptable. Low accuracy is usually related to underfitting and this motivated our decision to enlarge our dataset using data augmentation techniques. A network trained on the enlarged dataset achieved a frequency of 71.10 percent on the Canadian French Dataset, 79.07 percent on the Italian dataset, 73.89 percent on the North American Dataset and finally 70 percent on the Berlin Deutsche Dataset and therefore we concluded that a neural network trained on MFCC values of audio files was able to accurately predict. The MFCC parameters are thus an effective way to gauge the accuracy of the emotions when compared to each other.

The plots as shown in Figure 5, depicts means of the individual MFCC coefficients obtained for all the cultures for each point. Start from the first graph, it can be deduced that all the cultures express the emotion 'Happy' in their own ways and no emotion is like each other across any gender. The same could be said about the emotion 'Angry' as no 2 emotions are like each other. However, there is a small hint of Berlin Deutsche and the North American datasets being like each other regarding the 2 emotions mentioned. This could be seen via the MFCC plots wherein the 2 cultures intersect each other at 2 points. A welcome break is seen in the emotion 'Sad' wherein it can be said that for the gender 'Female', Italian and the North American emotions are in some ways similar as them.

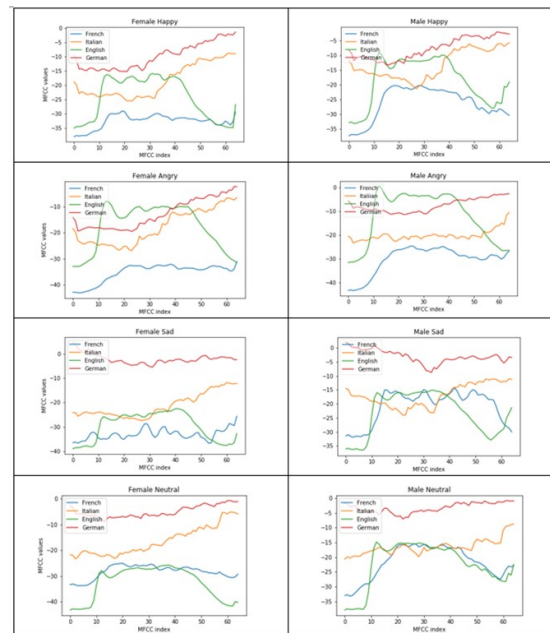


Fig.5. MFCC Plots

MFCC values overlap each other at certain points. Berlin Deutsche dataset although remains totally different to the other 3 cultures combined. For the same emotion, the gender 'Male' expresses likeliness across each dataset as there is an overlap at many points. Here again, Berlin Deutsche expresses the emotion in a different way. The emotion 'Neutral' again shows the same trend as seen in 'Sadness', but here for the gender 'Female' the 2 cultures showing similar traits are the Italian and the North American dataset. While for the gender 'Male', all cultures express the emotion 'Neutral' in the same way as to each other with the Berlin Deutsche following suit.

5 Discussion

Overall, For the Canadian French, Italian and the North American, emotions are expressed in a way that are recognizable across these cultures. So, an individual from a cultural background of any of these 3 should not have high difficulty in striking a conversation. Same however could not be said about the Berlin Deutsche Dataset. People from a German background would probably find it difficult to express their thoughts in a way that would lead to desired consequences. Having said that, further research needs to be carried across other cultures to see more significant comparisons and deduce as to whether there strikes a possibility of a group of cultures being like each other.

6 Acknowledgement

This work was partially supported by the Bulgarian Ministry of Education and Science under National Scientific Program “Information and Communication Technologies for a Single Digital Market in Science, Education and Security”, approved by DCM No 577, 17 August 2018.

References

1. Hareli, S., Kafestios, K., and Hess, U. : A cross-cultural study on emotion expression and the learning of social norms. *Journal of Frontiers in Psychology* (2015)
2. Gournay, P., Lahaie, O., and Lefebvre, R.: A canadian french emotional speech dataset. In: 9th ACM Multimedia Systems Conference, pp. 399–402. , Amsterdam (2018)
3. Costantini, G., Iaderola, I., Paoloni, A., and Todisco, M: Emovo Corpus: an Italian Emotional Speech Database In: International Conference on Language Resources and Evaluation, pp. 3501–3504. Association of Computing Machinery (2014)
4. Livingstone, S.R., and Russo, F.: Livingstone, S.R., and Russo, F. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *Journal of PLoS ONE* (2018)
5. Burkhardt, Felix and Paeschke, Astrid and Rolfes, M. and Sendlmeier, Walter and Weiss, Benjamin. : A database of German emotional speech In: 9th European Conference on Speech Communication and Technology, pp. 1517–1520. Interspeech'2005 - Eurospeech (2005)
6. Rebai, I., BenAYED, Y., Mahdi, W., and Lorre, J.-P. : Improving Speech Recognition using Data Augmentation and Acoustic Model fusion. *Journal of Procedia Computer Science* (2017)
7. Iliev, A. I., and Stanchev, P. L. : Glottal Attributes Extracted from Speech with Application to Emotion Driven Smart Systems In: Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management , pp. 297–302. Seville, Spain (2018)
8. Iliev, A. I. : Emotion Recognition From Speech. Lambert Academic Publishing (2012)
9. S. Albawi, T. A. Mohammed and S. Al-Zawi: Understanding of a convolutional neural network In: 2017 International Conference on Engineering and Technology (ICET), pp. 1–6. , Antalya (2017)
10. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from., *Journal of Machine Learning Research* (2014).
11. Hue, A. , Dense or Convolutional Neural Network. Medium.
12. Alexander Iliev, Ameya Mote, Arjun Manoharan : Cross-Cultural Emotion Recognition and Comparison Using Convolutional Neural Networks In: Cross-Cultural Emotion Recognition and Comparison Using Convolutional Neural Networks”, Digital Presentation and Preservation of Cultural and Scientific Heritage Conference Proceedings, Vol.10. Sofia, Bulgaria: Institute of Mathematics and Informatics (2020)